APLICACIÓN DE MÉTODOS DE INTELIGENCIA ARTIFICIAL EN EL ÁREA MÉDICA

Norma Verónica Ramírez Pérez
Instituto Tecnológico de Celaya
norma.ramirez@itcelaya.edu.mx

Martín Laguna Estrada
Instituto Tecnológico de Celaya
martin.laguna@itcelaya.edu.mx

Ana Miriam De Anda Tovar
Instituto Tecnológico de elaya
miriam.deanda@gmail.com

Resumen

El presente artículo muestra los resultados obtenidos de un estudio realizado en el área médica, específicamente de patologías presentes en la columna vertebral (Enrique da Mota, 2011), para este estudio se dispuso de una base de datos extraida del repository UCI, http://archive.ics.uci.edu/ml/. El análisis de la información fue realizado con el software WEKA utilizando algoritmos de clasificación como LMT, Redes Bayesianas, Naive Bayes, J48 y Naive Bayes Simple. Se presenta al final de este trabajo una comparación del funcionamiento de estos clasificadores para decidir cuál es el mejor de ellos en el diagnóstico de patologías en la columna vertebral.

Palabra(s) Clave(s): J48, LMT, Redes Bayesianas, Naive Bayes, WEKA.

1. Introducción

En la literatura existen artículos científicos sobre la aplicación de técnicas de aprendizaje automático en la medicina, algunos estudios con respecto a este

problema describen varias aplicaciones en diferentes campos como la cardiología, gastroenterología, neumología, oncología, neurología, otorrinolaringología, ginecología y obstetricia, oftalmología, radiología, patología, citología, genética y bioquímica entre otros.

En el área de medicina, se han hecho investigaciones sobre la columna vertebral, la cual es un sistema compuesto por un grupo de vértebras, de discos invertebrados, nervios, músculos, médula y articulaciones (Henry Rouvière 1998). Las principales funciones de la columna vertebral son las siguientes: eje de soporte del cuerpo humano, protector óseo de la columna vertebral, la médula y raíces nerviosas. Es base de referencia de los ejes de movimiento del cuerpo, lo que hace posible el movimiento en tres niveles: frontal, sagital y transversal. Este sistema complejo desgraciadamente puede sufrir disfunciones que causan dolores de espalda con muy diferentes intensidades. Hernia de disco y espondilolistesis son ejemplos de patologías de la columna vertebral que causan dolor intenso y son el resultado de traumas –con frecuencia nuy pequeños-, en la columna que van dañando progresivamente la estructura del disco intervertebral.

Para llevar a cabo este estudio sobre diagnóstico de patologías en la columna vertebral, se utilizaron algoritmos que pertenecen a dos tipos principales de modelos clasificadores: el primero versa sobre la base de los árboles de decisión (Lior Rokach and Oded Maimon, 2008) y el segundo, sobre los clasificadores bayesianos (Lior Rokach and Oded Maimon, 2008). Cabe mencionar que estos algoritmos se trabajaron con el software libre WEKA, desarrollado en la Universidad de Waikato, Nueva Zelanda, considerado como una plataforma de aprendizaje automático.

Un árbol de decisión (Covert, 1967) es una estructura de árbol diagrama de flujo similar o modelo de decisiones, donde cada nodo interno denota una prueba en un atributo, cada rama representa un resultado de la prueba que conduce a un nodo hoja en representación de clases o distribuciones de clase. El nodo superior en un árbol es el nodo raíz. Los árboles de decisión se construyen de manera recursiva divide y vencerás de arriba hacia abajo. A partir de un conjunto de entrenamiento de tuplas y sus etiquetas de clase asociados, el conjunto de entrenamiento se divide de forma

recursiva en subconjuntos más pequeños a medida que se construye el árbol, sin embargo no todas las ramas se ven en un árbol de decisión. Una acción que permite mejorar la precisión en la clasificación es la poda de árboles, la cual intenta identificar y eliminar las ramas que pueden reflejar el ruido o los valores extremos con el objetivo mencionado de mejorar la precisión de la clasificación.

2. Conceptos preliminares

Los algoritmos dedicados al problema de la clasificación supervisada operan usualmente sobre la información suministrada por un conjunto de muestras, patrones, ejemplos o prototipos de entrenamiento que son asumidos como representantes de las clases y poseen o se les asigna una etiqueta de clase correcta. A este conjunto de prototipos correctamente etiquetados se le llama conjunto de entrenamiento (TS, training set), y es el conocimiento empleado para la clasificación de nuevas muestras. Estos algoritmos tienen como objetivo determinar cuál es la clase de las que ya se tiene conocimiento a la que debe pertenecer una nueva muestra, teniendo en cuenta la información que se puede extraer del conjunto de entrenamiento.

Los algoritmos utilizados en este estudio son LMT(Landwehr, 1985), Redes bayesianas(Mitchell,1997), Naive Bayes(Mitchell,1997), J48 (Ross Quinlan, 1993), Naive Bayes simple(Mitchell,1997). Las redes bayesianas organizan los datos mediante un conjunto de varables y analizan la relación entre ellas, es decir, estiman la probabilidad de las variables no analizadas en base a las variables analizadas. Por otro lado, el algoritmo Naive Bayes, predice la probabilidad de los posibles resultados, y se utilizó en este estudio para realizar la exploración inicial de los datos. Por otra parte, el algoritmo Naive Bayes simple utiliza las probabilidades de cada variable para hacer una predicción, en este algoritmo los atributos numéricos se modelan mediante una distribución normal. Finalmente el algoritmo J48 construye un árbol a partir de los datos realizando nodos que minimicen la diferencia entre los datos utilizando atributos numéricos para generar el árbol.

3. Metodología

Para realizar este estudio se utilizó la base de datos "Vertebral Column" la cual fue extraída de Uci Repository, plataforma utilizada en el ámbito académico para realizar pruebas con las bases de datos de donación, donde los investigadores comparten sus investigaciones para que otros puedan hacer uso de ellas. Esta base de datos mencionada hace una clasificación de los pacientes en tres categorías: normal, con hernia discal y con espondilolistesis.

A continuación se describe en forma breve la información de los atributos. Cada paciente está representado en el conjunto de datos por seis atributos biomecánicos derivados de la forma y orientación de la pelvis y la columna vertebral lumbar (en este orden): incidencia de la pelvis, inclinación de la pelvis, ángulo lordosis lumbar, sacra de pendiente, el radio de la pelvis y grado de espondilolistesis. La siguiente convención se utiliza para las etiquetas de clase: hernia de disco (DH), espondilolistesis (SL), Normal (NO) y anormal (AB).

Cabe mencionar que para poder utilizar la base de datos no fue necesario normalizar los datos ya que al momento de descargar la base de datos Uci Repository, proporcionó un archivo para el manejo de los datos en entorno WEKA.

4. Resultados

Los resultados obtenidos durante la clasificación de los datos por medio de los algoritmos utilizados muestran que el algoritmo de decisión LMT clasificó 265 instancias correctamente y solo 45 instancias incorrectamente de un total de 310 instancias con un porcentaje de éxito de 85.483%, y un error absoluto de 0.2168. El algoritmo Redes Bayesianas clasificó correctamente 237 de las 310 instancias con un porcentaje de éxito de 80% y un error absoluto de 0.249; el algoritmo Naive Bayes clasificó correctamente 248 instancias con un error absoluto de 0.2, mientras que J48 clasificó 253 instancias, y finalmente el algoritmo Naive Bayes Simple clasificó correctamente 241 instancias de las 310. Como observamos en la tabla 1 se puede verificar que el mejor algoritmo de decisión es el LMT mientras que el mejor clasificador bayesiano es Naive Bayes.

Tabla 1 Comparación de algoritmos de clasificación.

Algoritmo	Instancias clasificadas correctamente	Instancias clasificadas incorrectamente	Error absoluto	Error relativo absoluto	Número total de instancias	Estadístico kappa	Error cuadrático relativo	% de Éxito	% de Error
LMT	265	45	0.2168	49.56%	310	0.6738	68.68%	85.48%	14.51%
Redes bayesianas	237	73	0.2409	55.06%	310	0.5134	90.54%	80%	20%
Naïve Bayes	253	62	0.2	50.31%	310	0.5621	92.37%	78.74%	21.26%
J48	248	57	0.2057	47.01%	310	0.5495	80.04%	78%	22%
Naive Bayes Simple	241	69	0.2201	45.71%	310	0.5023	92.30%	77.74%	22.26%

En la tabla 1 observamos que el estadístico kappa corresponde a la proporción de instancias observadas sobre el total de instancias de la base de datos, este estadístico toma valores entre -1 y +1. Cuando los valores son cercanos a 1 significa que es mayor el grado de relación entre los datos, cuando el valor es cero significa que los valores son los esperados ya que las instancias fueron tomadas al azar, por lo que se concluye en la tabla1 que este estadístico muestra un 0.6738 en el algoritmo LMT, es el más cercano a 1 por lo que su grado de relación entre los datos es el más elevado y por ende es el que mejor clasifica. En la figura 1 se muestra la clasificación de los algoritmos en donde se destaca que el algoritmo con el mejor porcentaje de éxito es el LMT con un 85.48 %.

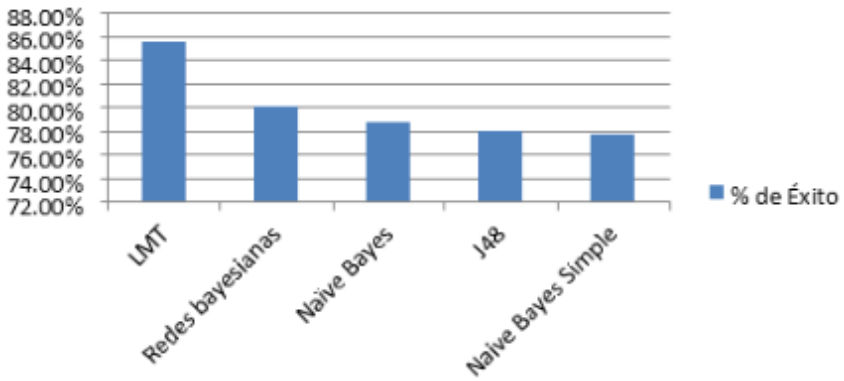

Figura 1 Porcentaje de éxito.

En la figura 2 podemos observar el algoritmo que clasificó correctamente la mayor parte de los datos con 265 instancias de un total de 310.

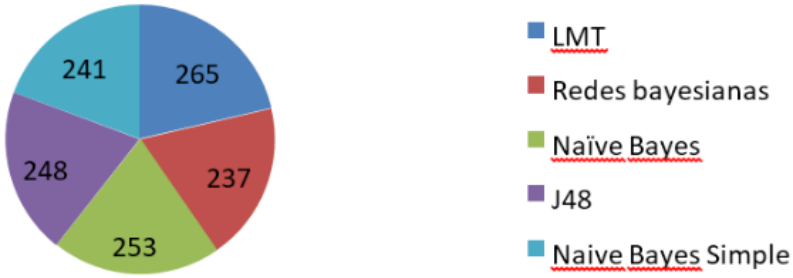

Figura 2 Instancias clasificadas correctamente.

4. Conclusión

El análisis de datos en este trabajo se llevó a cabo con WEKA la cual es una herramienta de aprendizaje que permite elegir los mejores algoritmos de minería de datos que se aplicaron sobre los datos seleccionados. A partir de los resultados obtenidos en este trabajo para el diagnóstico de problemas en la columna vertebral, se llegó a la conclusión de que el mejor algoritmo de clasificación es Naive Bayes, que clasifica los casos con una precisión de 92.37%, mientras que el mejor algoritmo de decisión es el algoritmo LMT clasificando los datos con una exactitud de 68.68%. Por otro lado, en el artículo original se encontró que el mejor clasificador es el Multilayer Perceptron (Quinlan,1996) ya que es el que mejor clasifica los datos con una precisión de 95% y es el que mejor ayuda proporcionó en el diagnóstico de las patologías de la columna vertebral, además de que dicho algoritmo permite acortar el tiempo de entrenamiento de los datos sin afectar su eficiencia.

5. Bibliografía

[1] Arellano, H. P. (s.f.). ecured. http://www.ecured.cu/index.php /Algoritmos_de_clasificaci%C3%B3n_supervi sada.

[2] Brownlee, J. (s.f.). Machine Learning Mastery.

[3] Berthonnaud, E., Dimnet, J., Roussouly, P. & Labelle, H. (2005). 'Analysis of the sagittal balance of the spine and pelvis using shape and orientation parameters', Journal of Spinal Disorders & Techniques, 18(1):40, 47.

[4] Hernandez, C. C. (s.f.). Clasificador naive bayes. http://naivebayes.blogspot .mx/.

[5] Rocha Neto, A. R. & Barreto, G. A. (2009). 'On the Application of Ensembles of Classifiers to the Diagnosis of Pathologies of the Vertebral Column: A Comparative Analysis', IEEE Latin America Transactions, 7(4):487-496.

[6] Rocha Neto, A. R., Sousa, R., Barreto, G. A. & Cardoso, J. S. (2011). 'Diagnostic of Pathology on the Vertebral Column with Embedded Reject Option,Proceedings of the 5th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2011), Gran Canaria, Spain, Lecture Notes on Computer Science, vol. 6669, p. 588-595.

[7] Heaton Jeff. Introduction to Neural Networks for Java, 2nd Edition. St. Louis, Missouri : s.n., 2005. ISBN: 1604390085. 440p. http://archive.ics.uci.edu/ml/.